MOOD

Monitoring Outbreak events or Disease surveillance in a data science context

Project number: 874850

Horizon 2020

SC1-BHC-13-2019

Type of action: RIA

Deliverable D3.3

Title: Linking chain (processing algorithms and protocols) for disease and covariate datasets

Due date of deliverable: 30/06/2023

Revised due date of deliverable: 31/12/2023

Actual submission date: 19/12/2023

Start date of the project: January 1st, 2020 **Duration: 60 Months**

Leader of the Deliverable: INRAE

Dissemination Level	
PU Public	X
PP Restricted to other programme participants (including the Commission Services)	
RE Restricted to a group specified by the consortium (including the Commission Services)	
CO Confidential, only for members of the consortium (including the Commission Services)	

Project Information	
Project Acronym	MOOD
Project Full Title	MOnitoring Outbreak events for Disease surveillance in a data science context
Name of the funder	European Commission – H2020
Call Identifier	H2020-SC1-2019-Single-Stage-RTD
Topic of the call	SC1-BHC-13-2019 Mining big data for early detection of infectious disease threats driven by climate change and other factors
Grant Agreement Number	874850
Project Duration (YYYY/MM - YYYY/MM)	2020/01/01 – 2024/12/31
Project coordinator (Name of institution)	CIRAD
Cirad Project scientific leader (name, RU, e-mail)	Elena ARSEVSKA, ASTRE, elena.arsevska@cirad.fr
Project goals	MOOD aims at using state of the art data mining and data analytical techniques of disease data, Big data, and contextual data originating from multiple sources to improve detection, monitoring, and assessment of emerging infectious diseases (EID) in Europe. MOOD will establish a platform for mapping and assessment of epidemiological and genetic data in combination with environmental and socio-economic covariates in an integrated inter-sectorial, interdisciplinary, One health approach. More precisely, MOOD will develop: 1. The epidemic Intelligence community of practice to identify user needs of end-users i.e. national and international human and veterinary public health organizations; 2. Data mining methods for collecting and combining heterogeneous Big data; 3. A network of disease experts to define drivers of disease emergence; 4. Data analysis methods applied to the Big data to model disease emergence and spread; 5. Ready-to-use online platform tailored to the needs of the-end users and complimented with capacity building and network of disease experts to facilitate risk assessment of detected signals. MOOD outputs will be co-constructed with end-users at public health agencies to assure their routine use during and beyond project duration. They will be tested and fine-tuned on a set of air-borne, vector-borne, multiple-transmission route diseases, including anti-microbial resistance and disease X. Extensive interactions with end-users, studies into the barriers to data sharing, dissemination and training activities and monitoring of the impacts and innovations of MOOD outputs will support future sustainable use.
Key words	Infectious diseases, big data, epidemic intelligence, one health, impact, environmental changes, climate changes, user needs, socio-technical innovation
Project partners (Name of institutions)	CIRAD, ITM, FEM, ETH, INESC ID, ERGO, SIB, INSERM, ULB, KU LEUVEN, UM, SOTON, AVIA-GIS, MUNDIALIS, OPENGEOHUB, UOXF, ISS, THL, GERDAL, IPHS, ISCIII, ANSES, INRAE, ISID

Executive Summary

The present report details the data normalisation protocols and the processing algorithms that were developed within the MOOD project, to support the linking of the different data sources of diseases and co-variate data. These efforts cover data from official sources of structured disease records, data collected from unofficial sources such as news data, and covariates from the MOOD GeoNetwork.

Keywords

Data linking, Integration, Normalisation, Spatial Information

Document History

Date	Revision	Comment	Author/Editor	Affiliation
30/11/2023	V0.1	Initial version	Robin Engler, Bruno Martins, Mathieu Roche, Maguelonne Teisseire	SIB, INESC-ID, Cirad, INRAE
11/12/2023	V0.2	Review and inclusion of comments	Roberto Interdonato, Rémy Decoupes, Sarah Houben, Bruno Martins, Tom Matheussen	Cirad, INRAE, INESC-ID, Avia-Gis
19/12/2023	V0.3	Revised version	All	Cirad, INRAE, INESC-ID, Avia-Gis, mundialis, SIB

Table of Contents

1. Introduction

One of the objectives of the MOOD project is to provide researchers with access to standardised and open-source records of disease outbreak events, including data from official sources of structured disease records, data collected from unofficial sources such as news data and social media, and data from other existing monitoring platforms. These records should include the date, location, host, and pathogen species associated with the outbreak of a disease within the geographical boundaries of the MOOD project.

The MOOD project envisions to make links to these disease data, and also the tools to standardise them, available via its online project platform (https://mood-h2020.eu/data-and-covariates-access), so that they can be used by both the disease modellers within the MOOD project, and by the broader scientific community in general.

Data linking in the context of event-based surveillance involves establishing connections between disease-related information and non-disease data (e.g., environmental covariates), as well as between different sources of data (e.g., structured and unstructured disease occurrence events), creating a comprehensive view of events. To achieve the integration and interoperability of data in the MOOD platform, the effective normalisation of data across the spatial, temporal, and thematic dimensions is essential. Data normalisation also facilitates data re-usability according to the three dimensions: space, time, and theme. Data normalisation and linking is provided as a service by the MOOD platform (as output), envisioning the use by end-users as well as modellers.

The present report details the processing algorithms and protocols that were developed to support the linking of disease and co-variate data, as well as the linking between disease data from different data sources. As illustrated in Figure 1, specifically through the red arrows, we use the term *"data linking"* when the linking process is performed on the stored data (linked data as output), and the term *"metadata linking"* when the linking process is performed on the metadata (list of datasets as output).

The linking of disease data to environmental covariates depends on the proper normalisation of the thematic, temporal, and geospatial dimensions of the data. Indeed, if data is normalised by following compatible representation formats for encoding these dimensions, then a straightforward linking can be performed by joining the data with basis on equality/compatibility for the values associated to these attributes (i.e., once data normalisation has been achieved, the actual linking of the data is relatively straightforward, requiring joins on tables based on common columns, or joining tables and raster datasets based on common geospatial coordinates).

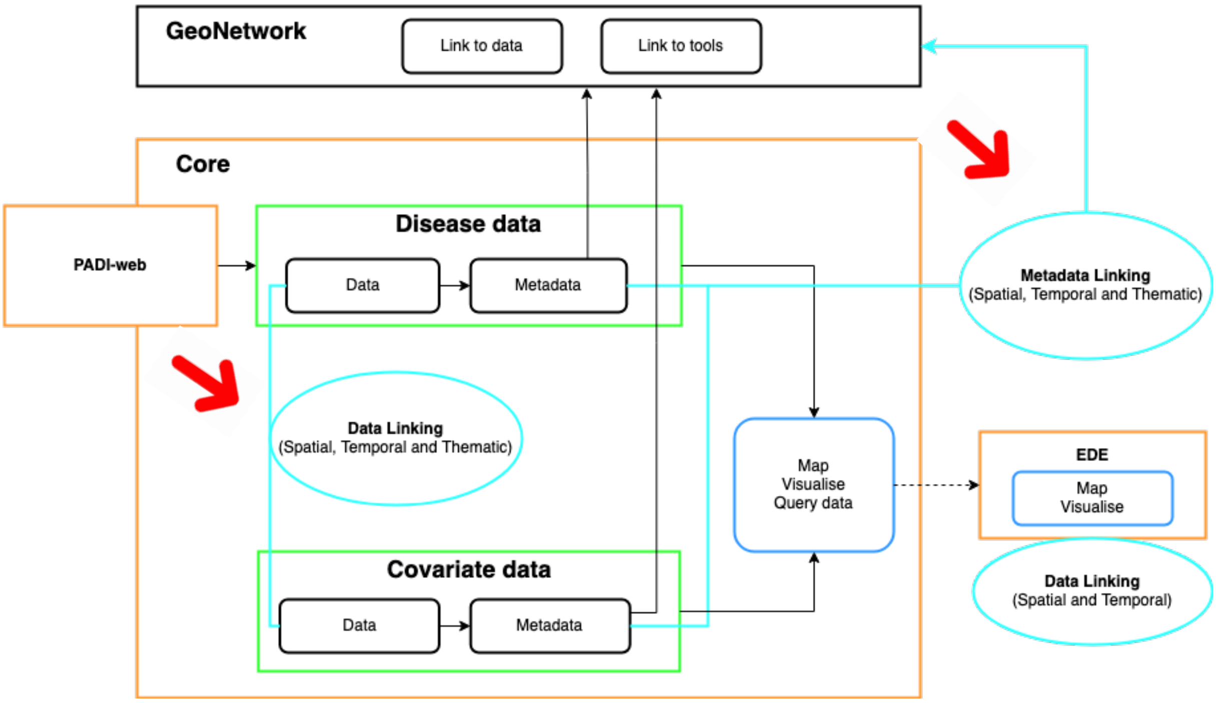

Figure 1: MOOD platform pipeline with the linking process.

2. The MOOD GeoNetwork

The entry point for prospective end-users of MOOD data is the MOOD platform (https://mood-platform.avia-gis.com/), which heavily relies on the MOOD GeoNetwork to store and catalog data.

The MOOD GeoNetwork (https://geonetwork.mood-h2020.eu/geonetwork) is a web-service that provides access to the metadata of all MOOD datasets and tools that are deemed useful for partners and prospective end-users. The GeoNetwork offers a range of information regarding these datasets. Various categories are available, including health, environmental data, elevation, meteorology, and climate, among others.

Currently, the MOOD GeoNetwork contains data and metadata for more than 100 disease data, 50 environmental variables, and 10 tools, that are browsable via a web-interface (Figure 2). A set of topics allows the user to index and query the catalog.

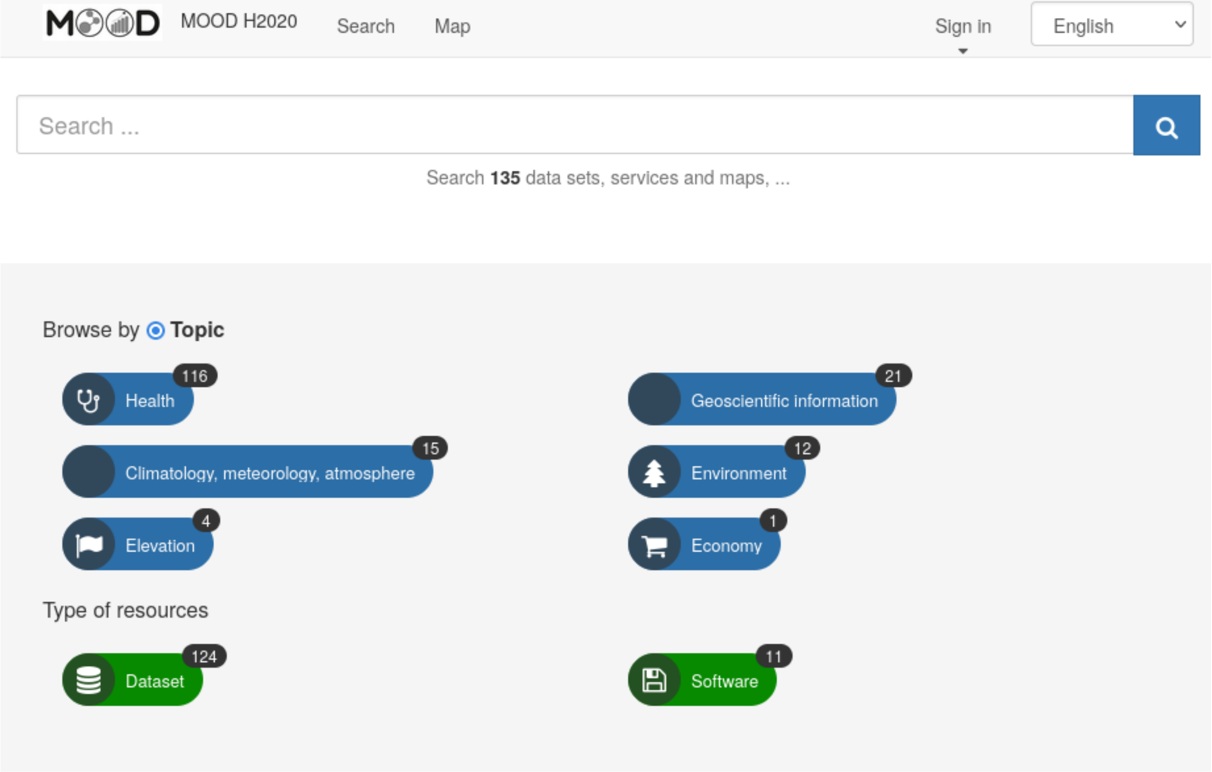

Figure 2 - The MOOD GeoNetwork landing page (Screenshot taken on 12.12.2023).

2.1 Populating the GeoNetwork

The population of the MOOD GeoNetwork, ingesting all the relevant information into the catalog, is currently performed using an R script (available at https://gitlab.irstea.fr/umr-tetis/mood/geonetwork-insertion). This script reads and parses an internal file of the MOOD project (i.e., the spreadsheet that is available at https://docs.google.com/spreadsheets/d/1_P01ZPObmbhMymaVDM547Rr2RIrw-gGX/edit#gid=699786557), which lists all MOOD productions. Subsequently, the script generates ISO-19115 records (a standard for describing metadata for geographic data). Following this procedure, a manual insertion of the ISO records is carried out. Additionally, further records are harvested regularly from the GeoNetwork of MOOD partners (e.g., https://data.mundialis.de/geonetwork/).

3. Normalisation of structured disease data

Structured disease data are records of outbreak events that are compiled and made available (with various degrees of open-accessibility) by governmental and non-governmental agencies. While such data are generally distributed in tabular format, the detailed formatting and units vary greatly, and are specific to each data source. The standardisation procedure must therefore be customised to fit each data source.

Within the MOOD project, the data normalisation pipeline (Figure 3) aimed at normalising outbreak records from the following three popular sources of structured disease data:

- **EMPRES-i** [https://empres-i.apps.fao.org]: The EMPRES-i (Global Animal Disease Information System) database is originally intended to support veterinary services by providing access to regional and global disease outbreak records. Its outbreak records include information on geographical location, date, pathogen and host species.
- **GenBank** [https://www.ncbi.nlm.nih.gov/genbank]: GenBank contains epidemiological outbreak records as well as genomic sequences. GenBank contains annotated collections of publicly available nucleotide sequences, along with isolated annotations that include coarse geographical location, date and host species.
- **ECDC** [https://www.ecdc.europa.eu/en]: The European Centre for Disease Control collects outbreak record data and provides them under certain conditions (not open-source data). Its outbreak records include information on geographical location, date, pathogen and host species.

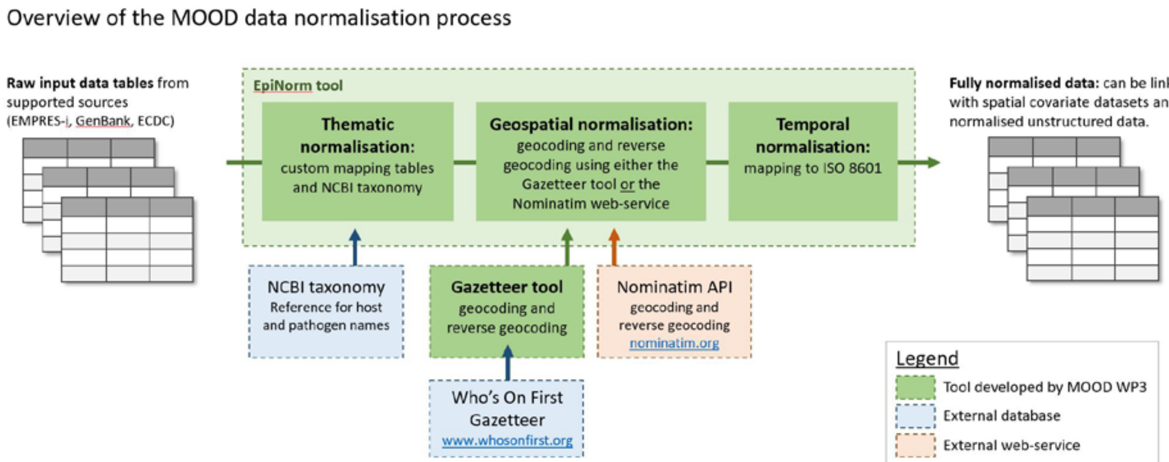
Overview of the MOOD data normalisation process
Raw input data tables from supported sources (EMPRES-i, GenBank, ECDC)
EpiNorm tool
Thematic normalisation: custom mapping tables and NCBI taxonomy
Geospatial normalisation: geocoding and reverse geocoding using either the Gazetteer tool or the Nominatim web-service
Temporal normalisation: mapping to ISO 8601
Fully normalised data: can be linked with spatial covariate datasets and normalised unstructured data.
NCBI taxonomy Reference for host and pathogen names
Gazetteer tool geocoding and reverse geocoding
Nominatim API geocoding and reverse geocoding nominatim.org
Who's On First Gazetteer www.whosonfirst.org
Legend
Tool developed by MOOD WP3
External database
External web-service

Figure 3: Overview of the MOOD data normalisation process.

The data normalisation pipeline is composed of two custom tools developed within WP3, namely *EpiNorm* (available at https://github.com/sib-swiss/epinorm) and the gazetteer access tool (available at https://github.com/bgmartins/gazetteer-access). More specifically, EpiNorm performs the overall data normalisation workflow, focusing on the thematic dimension of the data, while the gazetteer access tool is used to perform the geocoding and reverse geocoding tasks that are critical to the normalisation of the data's geospatial dimension (see Figure 3). The EpiNorm tool is described in more detail in milestone M34, while the Gazetteer tool is in-turn described in milestone M32.

In the next sections, we discuss the normalisation of the data in its three dimensions: thematic, spatial and temporal.

3.1 Normalisation on the thematic dimension

The thematic dimension of the data comprises both the host and the pathogen species (or disease) names. The normalisation of host and pathogen species names in outbreak records is important because not all data sources use the same vocabulary. For instance, some sources might refer to infected hosts using common names, while others use binomial nomenclature (i.e. scientific names).

To normalise data in the thematic dimension, we convert all host and pathogen species to the NCBI reference taxonomy (Federhen, 2012), using mapping tables developed within the project. The NCBI taxonomy was chosen because of its wide user base and exhaustivity.

In addition, we also perform pathogen serotype extraction. Specifically, the serotype of the pathogen involved in the outbreak is extracted from the pathogen name provided in the original outbreak record.

3.2 Normalisation on the temporal dimension

Handling the temporal dimension of the data is relatively straightforward, requiring the normalisation of dates expressed in textual/numerical form into a canonical format. In our case, we follow the ISO-8601 format (YYYY-MM-DD). The practical implementation of the normalisation relies on an existing Python library named dateparser[1] to interpret human readable dates in a variety of different formats, converting the temporal references into the ISO-8601 calendar date format.

3.3 Normalisation on the spatial dimension

Normalising the geospatial dimension of the data is particularly challenging, given the need to handle different representation formats (e.g., raster files associating counts to cells within a regular tessellation of the geographic space, vector files associating counts to polygonal boundaries, or tabular data in which one of the attributes uses place names to refer to geographic locations) and different granularities/resolutions for the geographic references (e.g., polygonal boundaries that can refer to countries, districts, etc.).

In our pipeline, the normalisation of the geo-referencing information associated with outbreak records is performed by retrieving, for each record, the latitude/longitude coordinates (WGS84) and the geographical administrative units associated with the outbreak location at three spatial scales: country level, administrative level 1 (ISO-3166-2), and populated place (e.g. city, town or village).

Initially, the tasks of geocoding (retrieval of coordinates associated with an address) and reverse geocoding (finding the administrative units associated with a coordinate location) were performed using an external web-service, specifically the Nominatim API. Nominatim is a publicly available web-service that allows its users to query OpenStreetMap data by name and address, and to convert geospatial coordinates of latitude and longitude into synthetic addresses (https://nominatim.org). However, this option has problems in terms of limitations associated with the use of the API. For instance, there are limits on the number of requests that can be made per minute, and as a result processing a large number of records can be somewhat slow. Furthermore, this service is also not allowing us to consider canonical geospatial regions that were deemed as interesting for the project (e.g., the VectorNet regions).

We therefore decided to develop a specific software package - hereafter referred to as the *gazetteer access tool* - to perform geospatial data normalisation and integration. The gazetteer access tool can be used on its own (see milestone M32 for full details on this tool), but in the context of the MOOD data normalisation pipeline, we use it as a backend for the EpiNorm tool to perform geocoding and

[1] https://dateparser.readthedocs.io/en/latest/

reverse geocoding. The gazetteer access tool thus becomes a replacement (or alternative) for the Nominatim service (Figure 3).

The gazetteer access tool is publicly available on GitHub[2], and its main functionalities can be summarised as follows:

- A geocoding functionality takes place names as input and returns polygonal boundaries from the matching gazetteer entries, giving preference to highly populated places in the case of ambiguous matches. This functionality can be accessed through a simple Python interface, or through a command line tool that can take CSV or Excel files as input, in which one of the attributes corresponds to the place names.

- A reverse geocoding functionality uses point-in-polygon or polygon intersection queries over the gazetteer entries, returning place names for point/polygonal boundaries provided as input. This functionality can also be accessed through a Python interface, or through a command line tool.

- Zonal statistics (e.g., averages or sums over polygonal boundaries) can be computed from raster data with a basis on place names. The tool's geocoding (or reverse geocoding) functionality is first used to transform the place names (or latitude and longitude coordinates) into polygonal boundaries. Then, the rasterstats[3] Python package is used to compute the aggregated statistics associated with the polygonal boundaries. Again, this functionality can be accessed through a simple Python interface, or through a command line tool that takes as input a CSV/Excel file (where one of the columns corresponds to place names or to geospatial coordinates) and one GeoTIFF file, producing as output a new CSV/Excel file with the added zonal statistics.

The tool was developed with reusability and extensibility in mind, envisioning the support for different types of integration mechanisms and workflows (e.g., pipelines of command-line tools, or integration through Python code using the provided Python interface). More information about the functionalities is available from the tool's GitHub repository.

4. Normalisation of unstructured disease data

Unstructured disease data are news articles collected by means of the Platform for Automated extraction of animal Disease Information from the web (PADI-web - https://www.padi-web-one-health.org), an automated bio surveillance system devoted to online news source monitoring for the detection of emerging/new animal infectious diseases. The collected news items are automatically classified as *"relevant"* or *"irrelevant"* using machine learning techniques. The relevant news corresponds to recent or current infectious animal health events.

Several strategies to extract events (e.g. strategies based on locations, relevant sentences, position on the information in texts, etc.) were proposed and integrated within PADI-web. Moreover, methods to improve classification tasks based on deep learning approaches (i.e., approaches based on the BERT

[2] https://github.com/bgmartins/gazetteer-access

[3] https://pythonhosted.org/rasterstats/

language model) and PADI-web datasets have been investigated (https://aclanthology.org/2022.lrec-1.399) [MOOD032] to be integrated into PADI-web.

We developed different algorithms to extract events from textual data (news):

- Events in relevant articles based on locations extracted with spaCy;
- Events in relevant articles based on locations extracted with spaCy model learnt specifically for this task with labelled data;
- Events at the beginning of the articles;
- Events in outbreak articles (i.e., document-based classification);
- Events in outbreak articles (i.e., document-based classification) and current event sentences (i.e., sentence-based classification).

These algorithms were integrated into PADI-web 3.0 (https://doi.org/10.1016/j.onehlt.2021.100357). All events are being normalised through the use of Geonames and dedicated dictionaries, although current efforts are investigating the possibility of also using EpiNorm and the *gazetteer access tool* for this task - see Section 3.

4.1 Normalisation on the thematic dimension

For unstructured data, the primary challenge prior to data normalisation (be it on the thematic, temporal or spatial dimension) is the actual identification/extraction of the event from a large corpus of unstructured text (typically news articles or social media feeds). Proper identification/extraction of an event is of critical importance, as the quality of the information extracted for a given event has a direct impact on our ability to normalise the event: e.g., extracting an incomplete species/pathogen name compromises the possibility of successfully mapping that species/pathogen against a reference taxonomy.

In this context, the MOOD WP3 team assembled a new dataset sourced from four event-based surveillance (EBS) Systems: ProMED, PADI-web, HealthMap, and MedISys. This dataset, meticulously annotated according to dedicated guidelines (https://doi.org/10.57745/MPNSPH), not only contributes to the overarching goal of normalising unstructured data, but also specifically aids in the thematic normalisation of content related to Antimicrobial Resistance (AMR) events and issues. This new dataset was developed during the MOOD AMR Hackathon (22.06.2022, Montpellier, France, https://mood-h2020.eu/mood-hack-antimicrobial-resistance-hackathon/).

After successful identification/extraction of an event, we normalise the host and pathogen/disease names by mapping them against the NCBI (National Center for Biotechnology Information) reference taxonomy (Federhen, 2012). This allows linking and interoperability with structured disease data.

4.2 Normalisation on the temporal dimension

The HeidelTime tool was adapted and integrated into PADI-web (https://www.padi-web-one-health.org) to extract **temporal information** in textual data. In the previous version of PADI-web, temporal information was extracted with spaCy without considering relative temporal information.

This temporal information is then normalised according to the ISO-8601 format (YYYY-MM-DD), following the same procedure as for structured data (see Section 3.2).

4.2 Normalisation on the spatial dimension

The **spatial information** is available in two forms: Absolute Spatial Information (ASI) corresponding to complete place names (e.g., *Paris*, *London*, or *Germany*), and Relative Spatial Information (RSI) combining place names and spatial relations/qualifiers (e.g., *south of Paris*, *north Madrid*, or *80 km from Rome*). It is particularly challenging to extract RSI from textual data and compute the corresponding geospatial region. We proposed algorithms and associated prototypes (https://doi.org/10.5194/agile-giss-3-16-2022) to address the following tasks: 1) extraction of relative spatial information from textual data, and 2) geotagging of this relative spatial information.

5. Data Linking

The linking process involves the linking of disease data to environmental covariates, and also the linking of disease data from different sources (e.g., from structured and unstructured sources). In both cases, the process relies on the proper normalisation of the thematic, temporal, and geospatial dimensions of the data. If the data are normalised into compatible representation units and formats for encoding these dimensions, their linking becomes straightforward and can be performed by joining the data with basis on equality/compatibility for the values associated to these attributes. The previous sections in this report have already summarised the relevant aspects related to the normalisation of the data, which results in normalised data that use the following representation formats.

- **Temporal dimension:** ISO-8601 calendar format (YYYY-MM-DD).
- **Spatial dimension:** latitude / longitude coordinates (WGS84), NUTS 2021 codes, or VectorNet regions (either through polygonal boundaries, or canonical names).
- **Thematic dimension:** NCBI reference taxonomy (Federhen, 2012).

5.1 Linking disease data and non-disease data

Linking of disease data (records of outbreak events) and non-disease data (typically environmental variables in a raster or vector format) is performed by 1) matching the data on their temporal dimension, and 2) spatially joining these data (i.e. matching the data on their spatial dimension).

The precondition and foremost aspect of data linking is therefore their normalisation on the temporal and spatial dimensions, which is described in detail in Sections 2, 3 and 4 of this document. Once the data are successfully normalised, their linking can be performed relatively easily, by virtually all software designed to handle geospatial data (i.e., performing a spatial join between points and polygons, or between points and raster layers, is a trivial task for any software made to handle geospatial data).

5.2 Linking disease data from different sources

In Deliverable D3.2, we proposed a framework to compare official and unofficial disease data. The methodologies at the core of this framework can be exploited to perform the linking between these two data sources. Specifically, we compared the events collected from official and unofficial sources in terms of two aspects: 1) spatio-temporal analysis (how the events are geographically and temporally distributed), and 2) thematic entity analysis (what thematic entities are extracted from the events and how they are related to spatio-temporal analysis). The proposed methods are publicly available online (https://gitlab.irstea.fr/umr-tetis/mood/compebs), and have been tested on a sample of data extracted from three major sources: PADI-Web, ProMED, and EMPRES-i (data available in Dataverse: https://doi.org/10.57745/Y3XROX and listed in Deliverable D2.2).

At a first stage, we deal with a task of event matching, i.e. the ability to automatically identify the same event when collected by two different sources. We propose an approximation scheme by modelling this task as an assignment problem, where we assess the similarity between each pair of events by comparing normalised hierarchical data.

Subsequently, in order to make a link on the spatial dimension of the events collected by different sources, we resort to geographic representativeness by taking the temporal aspect into account. Specifically, we refer to *"spatio-temporal representativeness"*, i.e. the ability to measure to what extent an unofficial source represents outbreak events that are known to have occurred within a given geographic zone (as reported by an official source) for a given time period. As a result, we obtain knowledge about which geographical zones are represented better by which unofficial sources.

Then, regarding the temporal dimension, the linking relies on two key concepts: the temporal event patterns and the timeliness. Concerning temporal event patterns, our objective is to capture the temporal event patterns for each source and then compare these results between official and unofficial sources. This includes taking into account aspects such as endemic diseases that can occur repeatedly in some regions during the year, and non-endemic ones that may occur occasionally based on surrounding circumstances. On what regards the timeliness, it can be defined as the ability of identifying disease events in a timeframe that enables the use of the information by decision makers. For each official event, we measure timeliness as the time difference between its official report date and the publication date of the same event in an unofficial source.

The last comparison stage, allowing us to link official and unofficial disease data sources, is the one dealing with the thematic dimension. The idea is to assess to what extent the information about an event as reported by official and unofficial sources brings the same quantity of information, e.g., if the event is described in its entirety of details. This analysis is performed by means of the visual analysis of hierarchical chord diagrams.

6. Conclusions

This report detailed the processing algorithms and protocols that were developed to support the linking of the different data sources of diseases and co-variate data in the MOOD project. All datasets, source code, and tools, are provided (or linked to) on the MOOD platform (see Deliverables D2.2 and D3.4).

Bibliography

- M. AlamSyed & al. GeoXTag: Relative spatial information extraction and tagging of unstructured text. 2022. In : 25th AGILE Conference on Geographic Information Science "Artificial Intelligence in the service of Geospatial Technologies''. Copernicus Publications, 1-10. (AGILE: GIScienceSeries, 3) AGILE Conference on Geographic Information Science (AGILE 2022), Vilnius, Lituanie, 14 Juin2022/17 Juin2022. https://doi.org/10.5194/agile-giss-3-16-2022

- J. Monteiro & al. A Co-Training Approach for Spatial Data Disaggregation. In: Proceedings of the ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, 2022. https://doi.org/10.1145/3557915.3561475
- Assembly of dataset including epidemiological parameters, virus variants, vaccines, and climate factors in Europe from August 2020 to October 2021 and built a Bayesian inference model to untangle the changing, real-world impact of NPIs and vaccination on European Covid-19 trajectories (Ge et al. Nature Communications 2022). The datasets and code are available at https://github.com/wxl1379457192/Vaccine-NPIs-in-EuropeV2.
- Nejat Arinik, Roberto Interdonato, Mathieu Roche, Maguelonne Teisseire: An Evaluation Framework for Comparing Epidemic Intelligence Systems. IEEE Access 11: 31880-31901 (2023) https://doi.org/10.1109/ACCESS.2023.3262462
- Federhen, S. (2012) The NCBI Taxonomy database. Nucleic Acids Research, 40, D136-D143.